Journal of Information Systems Engineering and Management
2025, 10(30s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Detecting Groundwater Quality with Optimizing Gradient Boosting Performance: The Role of Focal Loss in Tree-Based Residual Models

R. Siddthan[1], Dr.PM. Shanthi[2]
Research Scholar[1], Assistant Professor[2]
[1,2]Department of Computer Science, J.J.College of Arts & Science (Autonomous) (Affiliated to Bharathidasan University), Pudukkotai, Tamil Nadu, India.
Corresponding Author: sithan314@gmail.com
shantisuman28@gmail.com

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Globally, groundwater serves as a vital source for drinking water and agricultural purpose for billions of people. Assessing the quality is indispensable for sensing harmful contaminants, comprising nitrates, heavy metals and pathogens. The constant observing and evaluation are significant to decrease the risks associated to the ground water contamination and to provide safe drinking water for people. Manual methods for assessing groundwater quality involves physically collecting samples from boreholes or monitoring wells and exposing them to laboratory analysis. Though these manual methods provide comprehensive insights into quality, being resource-intensive and time-consuming by the frequency and number of sampling sites. To solve this issue, several researches have concentrated on water quality detection. Conversely, it results with lacks in accuracy and analysing in safety or not. For feature optimisation, PCA (Principal Component Analysis Optimization) Model is utilised. For enhancing the classification performance, the proposed method employs ML (Machine Learning) approach called Optimized Gradient Model with Effective analysis on trees with focal loss. It utilizes water quality dataset for evaluation. For improving accuracy in performance, the present research incorporates focal loss method which eradicates the imbalances in the dataset. Correspondingly, efficacy of the present model is calculated utilising various performance metrics are F1 Score, recall, precision and accuracy to estimate the performance. Further, the internal comparison of proposed models such as AdaBoost, XG Boost, Gradient Boosting and Random Forest (RF) and traditional models discloses the effectiveness of the present ML method. The projected research envisioned to contribute the emerging water quality detection models thereby, protecting public health and preserving safety environment.

Keywords: Ground Water, Efficient Analysis, Principal Component Analysis, Optimized Gradient Model, Machine Learning, Focal loss and Tree based Residual model.

1. INTRODUCTION

Basically, groundwater plays an important role in ecologies and is an vital resource for global regions, predominantly in arid areas [1]. Poor groundwater quality could lead to substantial economic consequences, like increased costs of treatment and reduced in agricultural production [2, 3]. Perceiving groundwater quality is vital for securing public health, preserving the environment, ensuring sustainable agricultural practices, and effectively managing water resources [4]. Through gaining insights into groundwater quality, authorities could make efficient decisions about water usage, treatment, and conservation strategies [5]. Moreover, traditional manual methods for assessing groundwater quality can often be costly and time-consuming [6]. Recently, researches have begun to use AI (Artificial Intelligence) technologies for water quality detection systems. Among these, ML (Machine learning) models provides more efficient and cost-effective alternative for calculating water quality parameters, minimizing the need for extensive laboratory testing [7]. These models can be adapted to many geographical contexts and scaled to accommodate various data inputs, making as versatile tools for evaluating groundwater quality across diverse

regions. The use of ML and AI in groundwater quality prediction [8] enhances the ability to analyze complex datasets and identify key factors influencing water quality, resulting in improved accuracy and reliability of predictions [9].

Consequently, several conventional researches have attempted to undertake water quality detection. For an instance, the existing model has employed the water quality detecting model using PCR (Principal Component Regression) technique. The WQI (Water Quality Index) [10] has calculated utilizing weighted arithmetic index technique. Then, PCA (Principal Component Analysis) has applied to Gulshan Lake-related dataset [11, 12] and dominant WQI parameters [13] have been extracted. Experimental results have demonstrated that the prevailing model has achieved 95% of prediction accuracy [14]. In the same way, the conventional research has concentrated on water quality classification utilizing ML models. It has been developed a real-time water quality monitoring system. The prevailing model has used ML classifiers such as RF (Random Forest), XG Boost (Extreme Gradient Boosting), SVM (Support Vector Machine), LR (Logistic Regression), DT (Decision Tree), MLP (Multi-Layer Perceptron) and CATBoost (Categorical Boosting) [15] on drinking water quality dataset. Among these classifiers, CATBoost has attained better performance and could be used in vast range of datasets [16]. Similarly, the classical model has aimed to create reliable and accurate ML models for the parameters on irrigation. ANN (Artificial Neural Network), LSTM (Long Short Term Memory) and MLR (Multi-Linear Regression) have employed to detect the water quality parameters for irrigation. The experimental values have shown that the ANN and MLR models have been attained accurate and better performance than LSTM model [17]. Likely, this existing method has focused on evaluating the performance of three machine learning algorithms: DNN, GBM (Gradient Boosting Machine), and XG Boost for assessing groundwater indices. The dataset utilized in this study was obtained from the India Water Resources Information System which has attained better performance [18].

To overcome these issues, proposed model uses certain set of procedures to improve the performances in detection through groundwater quality classification. The proposed model utilizes publicly available water quality dataset. Once the input data are loaded, it processes with secveral pre-processing techniques such as checking missing values, normalization and label encoding. For feature optimization process, the proposed method utilised PCA optimised model. Then, extracted data are divided into two in the ratio 80:20 where 80 percent for training process and 20 percent for testing process. The training set is passed to the proposed ML model of Optimized Gradient which has effective analysis on trees to perform the groundwater quality classification performance. Focal loss method is incorporated to enhance the efficiency of data to avoid imbalanced data in dataset. Then, the testing data is evaluated through performance metrics. The major contributions of the present ML model is provide below.

- To utilise ML model for groundwater quality prediction with water quality dataset to improve the computation and accuracy in the proposed ML model.
- To deploy PCA Optimization model for the process of feature optimization.
- To employ Optimized Gradient Model with focal loss to accomplish Groundwater quality classification in an effective manner.
- To evaluate the efficacy of the present system with performance metrics are F1 score, accuracy, precision and recall.

The flow of the present model is given here: section 1 provides overview on the background of the proposed model. Section 2 reviews the conventional literatures related to water quality detection and problem identification. Section 3 precisely describes about proposed methodology. Furthermore, section 4 provides table and graphical representation of data analysis. Section 5 discuss the results of the proposed model with traditional researches. Finally, section 6 concludes the present model along with future researches.

2. LITERATURE REVIEW

This section provides about the analysis of various existing researches on water quality detection along with other techniques for the prediction on quality classification system.

For an instance, the prevailing research has developed a system for monitoring water quality on aqua culture ponds. It has based on the technology of NB-IoT (Narrow Band Internet of Things). Particularly, the designs and executions hardware and software like background control modules, terminal sensor nodes and monitoring applications and understands remote monitoring of ponds. An intellectual control of equipment like aerator are used for the existing system. It has maintained the temperature control accuracy with ± 0.12℃ [19]. Similarly, the existing model has presented an IoT based water quality monitoring network which continuously checks and evaluates the water quality.

It has tried to differentiate whether it is appropriate for general usage or not. The prevailing model has included with several sensors usages which measures different parameters of the water quality. The parameters such as pH, turbidity, temperature, conductivity and DO (Dissolved Oxygen). It has attained better performance [20]. Contrarily, the considered model has combined CNN (Convolutional Neural Network) and LSTM for stimulating the level of water and 3 water quality parameters such as TN (Total Nitrogen), TP (Total Phosphorus) and TOC (Total Organic Carbon) in NRB (Nakdong River Basin). Among these DL models, CNN has adopted for stimulating water level and LSTM has been used for stimulating the pollutants concentration. It has used RTWQI (Real-Time Water Quality Information) which has attained better performance [21].

Concomitantly, the traditional method has deployed a combined 2 DL models CNN and LSTM for predicting the variables of water quality such as Chl-a (Chlorophyll-a) (lg/L) and DO (Dissolved Oxygen) (mg/L) in Small Prespa Lake which has located in Greece. Results have shown that the conventional LSTM has outperformed than CNN for prediction [22]. Correspondingly, the recommended model has designed a WQM (Water Quality Monitoring) for monitoring drinking water quality that made use of IoT technology. It consisted of various sensors to evaluate different parameters like water turbidity, pH value, and water level in tank, humidity and temperature of the environment. Results have shown that the measured pH values are ranged from 6.5 to 7.5 and 7 to 8.5 whereas the turbidity has ranged from 600 to 2000 NTU (Nephelometric Turbidity Unit) for Metropolitan city supply water and groundwater respectively [23]. Contrastingly, the conventional model has aimed at forecasting the PS (Potential Salinity), RSC (Residual Sodium Carbonate), TDS (Total Dissolved Solid), SAR (Sodium Adsorption Ratio), MAR (Magnesium Adsorption Ratio) through T (Temperature), EC (Electrical Conductivity) and pH as input values. The prevailing model has used SVR (Support Vector Regression), AdaBoost (Adaptive Boosting), and RF and ANN models. Among these RF and AdaBoost have attained better performance than ANN and SVR [24].

Similarly, in the conventional model, ANFIS (Adaptive Neuro-Fuzzy Inference System) model has been employed to detect the WQI. FFNN (Feed-Forward Neural Network) and KNN (K-nearest neighbors) have been utilised to identify water quality. ANFIS model has shown accuracy of 96.17% for detecting WQI. It has utilized Indian water quality data [25]. Congruently, the purpose of existing model has been to utilise ML techniques like RF, MLR, NN, and SVM to classify the dataset. It has employed Indian water quality data. Experimental findings have shown better performance [26]. Likely, the traditional method has employed 13 physical and chemical parameters of water quality and 7 ML models, including DT, Gradient Boosting, RF, ANN, KNN, Naïve Bayes, and SVM. The ensemble model of Gradient Boosting with a learning rate of 0.1 has exhibited the better prediction performance compared to other models. It has achieved the better accuracy of 94.90%, sensitivity of 80.00%, and F-measure of 86.49%, with the lowest classification error [27]. Likewise, in the prevailing research, a deep learning-based Bi-LSTM model DLBL-WQA has been developed to forecast the factors of water quality in Yamuna River, India. The performance of prevailing model has been compared with various state-of-the-art techniques such as SVR, RF, ANN, LSTM, and CNN-LSTM. Experimental analysis has shown through calculating the BOD and COD levels, and has attained better performance [28].

2.1. Problem Identification

This section describes that several conventional methods have been limited by predicting the water quality which has several lacks.

- The conventional research needs to be carried out other important parameters such as residual chlorine, and nitrates in the water [24].
- Accuracy is a significant metric to measure the efficiency of the model. However, several existing researches lacks in providing accuracy [14] [25] [27].
- Many conventional researches have not focused on various chemical compounds in the water to detect its quality [7] [17, 27].

3. PROPOSED METHODOLOGY

The proposed method identifies and extract information from the given dataset for ground water quality detection. The detection is carried out by implementing Optimized Gradient, a machine learning model with focal loss. Existing works on the water quality detection have produced inaccurate results with slow speed convergence. Therefore, the proposed model utilised the ML algorithm for detection of water quality data. Moreover, flow of the proposed ML method is depicted in below figure. 1.

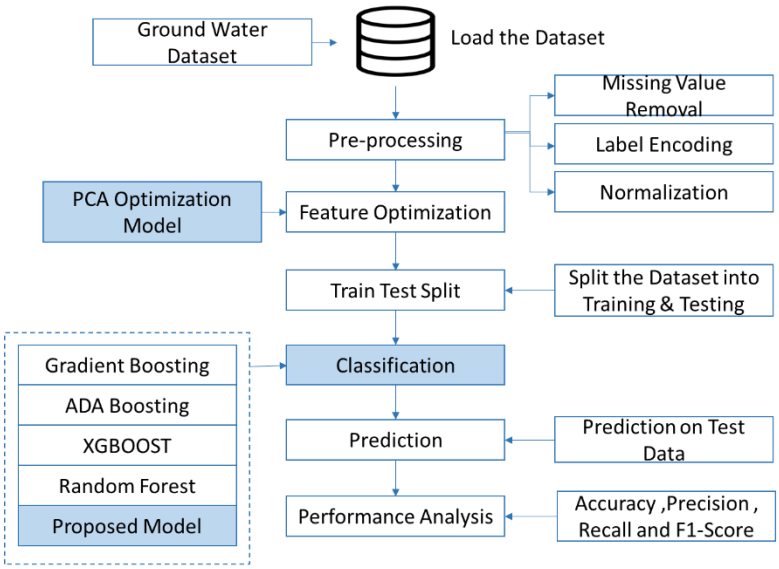

Figure. 1 Overall Design of the Proposed ML Model

The figure. 1 deliberates the proposed ML model's flow. It signifies the proposed system creation on ground water quality classification. It comprises of dataset loading, data pre-processing, and feature optimisation with the help of PCA optimization model, data splitting and classification process using Gradient Boosting, AdaBoost, XG Boost, Random Forest and proposed Optimized Gradient model. Following, the figure. 2 shows the mechanism of proposed method. The figure. 2 signifies the architecture of Present Ml model.

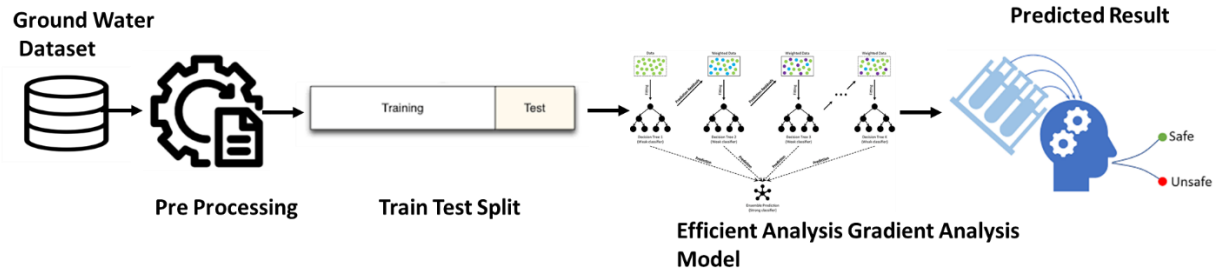

Figure. 2 Architecture of Present Optimized Gradient Model

Above Figure. 2 shows the proposed ML model's architecture. Once the input data is loaded to the model, it further pre-processes for retrieving clear data using techniques of checking missing values, normalization and label encoding. It improves the feature extraction process with accurate data that splits into two for training and testing purposes. The train set data is processed with efficient analysis Gradient Model to predict the water quality results with safe and unsafe classifications.

3.1. Dataset Description

The proposed Ml model uses water quality dataset which is based on replicated water quality measurements in an urban environment. It includes with numerous parameters with specified thresholds including safe and unsafe levels. A water quality dataset is designed to assess water potability and suitability for human usage for drinking. The dataset also includes a class attribute called "is_safe", indicates whether the water is safe (1) or not safe (0) for consumption based on the measured parameters. The official dataset link is provided below.

Dataset Link: ***https://www.kaggle.com/datasets/mssmartypants/water-quality***

3.2. Data Pre-processing

Data Pre-processing is a method of changing the raw data into proper data set which is processed to check the missing values, label encoding, feature scaling and data normalization, before applied to the algorithm. Besides, the pre-processing improves the feature optimization and classification performance of the proposed method. To achieve

this, proposed model executed three significant pre-processing techniques like checking missing values, normalization and label encoding. In the process of label encoding, the categorical data values are assigned to numerical labels to enhance the efficiency of the proposed model.

3.3. Feature Optimization- Principal Component Analysis Optimized Model

The present model used the feature extraction process for selecting the useful and most relevant data from the dataset to enhance the performance of the ML model. This methods are aimed to improve the accuracy, interpretability and efficiency through focusing on relevant input data. For feature optimization process, the proposed method utilised Principal Component Analysis Optimized Model. PCA converts a collection of correlated variables into a collection of uncorrelated variables referred to as principal components. This conversion is accomplished through an orthogonal transformation that maximizes variance, enabling the discovery of patterns within the data. The principal components are linear combinations of the original variables, arranged in a way that the initial components capture the majority of the variation found in the dataset.

3.4. Data Splitting

In Machine learning, this technique is used to eliminate the data over fitting issue. Essentially, ML uses the data splitting method to train the respective research where the train data is given to the proposed research for equipping the training stage parameters. After the training process, the test set data are measured to calculate the present model for handling the observations. In the present model, the original data is split into 2 sets in the ratio 80:20. The eighty percent of the new observations is utilised for the training and the remaining 20 percent of the data are applied for testing process to calculate the performance of the present research.

3.5. Comparison Algorithms

3.5.1 AdaBoost

Adaptive Boosting, known as AdaBoost, a robust ensemble learning method which has the potential to enhance the performance of classifiers substantially when it comes to detecting groundwater quality. During training phase, AdaBoost employs a weak learning algorithm sequentially, commonly decision trees, on training data. Each new classifier concentrates more on the misclassified instances through the previous classifiers. It is accomplished by modifying the weights of training instances, increasing the weights for misclassified instances and decreasing them for correctly classified. The present approach enables AdaBoost to attain a higher accuracy than any single weak classifier.

Pseudo code for AdaBoost

$Input:\ X: Ground\ Water\ Dataset;\ Y: Target\ Dataset;$
$T: Number\ of\ steps;$
$Output: H = a, the\ final\ hypothesis$
$Create\ D_0\ initial\ distribution$
$for\ t = 1\ to\ T\ do$
$Compute\ b_{t+1}\ from\ b_t;$
$Construct\ a\ a_t\ classifier\ from\ b_t;$
$end\ for$
$for\ (all\ x_i \in S\ test\ records)\ do$
$a(x_i) = a_1(x_1), a_2(x_2) \ldots, a_m(x_m);$
$end\ for$

3.5.2 Gradient Boosting

Gradient Boosting, a robust machine learning method which demonstrated an important effectiveness in detecting and classifying the groundwater quality. As an ensemble method, it combines multiple weak models particularly, decision trees to form a strong predictive model. Furthermore, the process involves iteratively adding new models which focuses on rectifying the errors of the preceding models. At an each iteration, Gradient Boosting minimizes a loss function, like cross-entropy for classification tasks or mean squared error for regression tasks.

Pseudo code for Gradient Boosting

Input: Take training set (a_i, b_i)
loss function L(y, T)
learning rate α ($0 < \alpha < 1$)
1: Make an initial simple model to classify data
2: $T_0(\mathrm{a}) = \arg min_\gamma \sum \mathrm{L}(b_i, \gamma)$
3: choose the round number M
4: for m $= 1, \ldots, \mathrm{M}$
5: *Measure* pseudo residual
6: $r_{im} = -\left[\frac{\partial \mathrm{L}(b_i, \mathrm{T}(b_i))}{\partial \mathrm{T}(b_i)}\right] T(b) = T_{m-1}(b)$
$i = 1, \ldots N$
7: Fit a base regression tree h_m to rim using (a_i, r_i)
8: for j $= 1, \ldots, J_m$, compute the multiplier
9: $\gamma_{jm} = \arg min_\gamma \sum \mathrm{L}(y, T_m(a_i) - \gamma_{hm}(\mathrm{a}))$
10: *end for*
11: *end for*
12: $Update\ T_m(a) = T_{m-1}(a) + \alpha.\ \gamma_m.\ \mathrm{b}_m(a)$
$Output$: $T_M(a) = T_0(a) + \sum_{m=1}^{M} \alpha.\ \gamma_m.\ \mathrm{b}_m(a)$

3.5.3 XG Boost

XG Boost is an advanced machine learning algorithm commonly used for classifying groundwater quality, thanks to its exceptional accuracy and efficiency. Its effectiveness in this domain stems from its capability to manage high-dimensional data and pinpoint key features that affect water quality. For instance, parameters such as EC (Electrical Conductivity) have been identified as critical for predicting water quality indices, whereas pH has been considered less influential.

Pseudo code for XG Boost

$Initialize\ q_0\ (x);$
$for\ n = 1,2 \ldots M\ do$
$measure\ g_n = \frac{\partial L(y,q)}{\partial q}$;
$measure\ h_n = \frac{\partial^2 L(y,q)}{\partial q^2}$;
$Determine\ \ by\ choosing\ splits\ with\ maximized\ gain$
$A = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} - \frac{G^2}{H}\right];$
$Determine\ the\ leaf\ weights\ w^* = -\frac{G}{H};$
$Determine\ the\ base\ learner\ \hat{b}(x) = \sum_{j=1}^{T} wI;$
$Add\ trees\ q_n(x) = q_{n-1}(x) + \hat{b}(x);$
end
$q(x) = \sum_{n=0}^{M} q_n(x)$

3.5.4 Random Forest

Random Forest is a highly effective machine learning method commonly utilized for classifying groundwater quality. Its versatility encompasses a range of water quality parameters, such as pH, turbidity, and hardness. This technique has been successfully applied in various studies to predict water potability and evaluate the ecological health of water bodies, highlighting its extensive applicability in environmental monitoring.

Pseudo code for Random Forest

Input: training data sets n_{N*P} *and Number of trees*(V)
For each variable $i \in P$ *do*
For $y = 1$ *to* V:
1. *Draw a sample* $B *$ *of size* O *from the training data.*
2. *Grow a random* $-$ *forest tree* F_y *to the* $\frac{2}{3}$ *of data.*
3. *Predict classification of the leftover* $\frac{1}{5}$ *using the tree, and calculate the classification rate = accuracy rate* (OOB)*, namely* $accuracy_y$.
4. *For variable* i*, permute the value of variable and compute the accuracy*$(accuracy_b)$*, subtract to the original oob error* $= accuracy_y - e_y)$*, the increase is an indication of the variable's importa*
End for

Aggregate total accuracy from all trees and calculate varia
$$\hat{\imath} = \frac{1}{A}\sum_{k=1}^{V} i_k \text{ and } s_i^2 = \frac{1}{V-1}\sum_{k=1}^{V}(i_k - \hat{\imath})^2$$
calculate importance of variable i: $e_i = \hat{\imath}/s_i$
End for

3.6. Classification - Optimizing Gradient Boosting with Focal Loss

The proposed model incorporates ML based model called proposed Optimized Gradient model to improve the classification results. The classification of the ground water quality in the respective model is processed with the proposed Optimized Gradient model which is trained in the ground water dataset. This section deliberates the details of equations and classification mechanism of the proposed method.

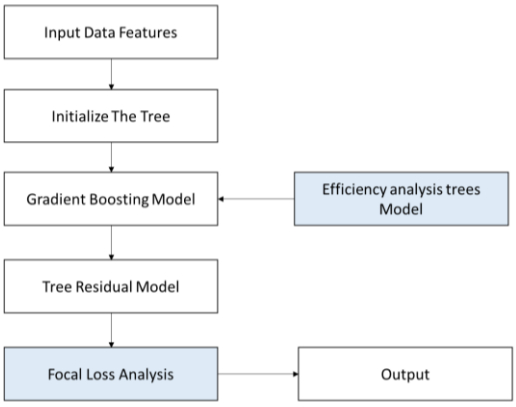

Figure. 3 Illustrative Diagram of Proposed ML Model

Figure. 3 describes an illustrative diagram of present Ml model. If the dataset is imbalanced, meaning certain classes are underrepresented, the initial predictions can be adjusted to account for the class distribution. For example, the model could be initialized to predict the majority class more frequently or use a weighted average based on class frequencies. In an optimized gradient boosting, the first step is to make an initial prediction for all instances in the dataset. This is commonly done using a simple model, such as predicting the mean of the target variable for regression problems or the log-odds for classification problems. This initial prediction serves as the starting point for the boosting process. The learning rate, also known as the shrinkage parameter, is crucial in this process. It scales the contribution of each tree, allowing for more controlled updates to the model. Using a smaller learning rate typically requires more trees to achieve optimal performance. To prevent over fitting, techniques such as limiting the depth of

the trees or using subsampling can be applied during tree initialization and subsequent iterations. This method enables the model to effectively learn complex patterns in the data. The pseudo code for proposed Optimised Gradient Model is given below.

Pseudo code for Proposed Model

$set\ p_0(w) = \left(\max_{j=1,\ldots,n}\{y_{1j}\}, \ldots \max_{j=1,\ldots,n}\{y_{sj}\}\right)$

For q = 1 to Q:

For j = 1, 2, . . . , n calculate $e_{jq} = y_j - p_{q-1}(w)$

Fit a deep efficient analysis to the targets e_{jq} given terminal regions R_{hq}, h = 1,2, . . . , J_q

For h = 1, 2, . . . , J_q calculate γhq = $d^{T(\aleph_q)}(R_{hq})$

Update $p_q(w) = p_{q-1}(w) + v \sum_{h=1}^{J_q} \gamma_{hq}\ \ I(w \in R_{hq})$

The proposed Optimized Gradient Model is seamlessly adapted from a single-output context to accommodate multiple outputs, it can also be demonstrated that each element of the vector $\hat{f}(x)$ adheres to the principle of monotonicity.

Focal loss

Focal loss is specifically designed to tackle class imbalance by reducing the loss contribution from easily classified examples while placing greater emphasis on those that are harder to classify. This approach is especially beneficial in water quality detection, where certain classes, such as "poor quality," may occur less frequently. During the training of each tree in the boosting process, focal loss can be utilized in place of traditional loss functions, such as mean squared error for regression or binary cross-entropy for classification. Once the model is initialized, each subsequent tree is trained to minimize the focal loss based on the residuals from prior predictions. This ensures that the trees concentrate more on misclassified instances, particularly those belonging to minority classes. The proposed Optimized gradient model utilizes Focal Loss for the classification of binary datasets, with the goal of reducing the differences between training and testing samples during the classification process, which often affects prediction accuracy. This adaptation of tree boosting greatly improves efficiency. It finds applications in multiple domains like water quality data analysis. With using the tree residual function, the focal loss analysis is used to enhance to a high extent. As a result, the Binary Focal Loss can be expressed by the following equation:

$$Z_{0_r} = -\sum_{m=1}^{p} log\ (\gamma_{hq})\ \zeta + \log(1 - p_m = 1\,\gamma_{hq})(1 - h_q)\,p_q(w) \quad (1)$$

When ζ is set to 0, the equation (1) above simplifies to the standard cross-entropy loss. To achieve the cross-entropy loss, the sigmoid activation function can be employed to improve accuracy and performance of the present model.

Correspondingly, tree initialization in the proposed optimised gradient boosting model with focal loss for water quality detection involves setting an appropriate initial prediction, using focal loss to address class imbalance, and iteratively training trees to improve predictions when focusing on misclassifying instances. The approach could enhance the proposed model's ability to detect and classify water quality levels accurately, in datasets where several classes are underrepresented. Bore wells are vital for domestic, industrial water supply, and agricultural uses. Key factors include water quality, pump efficiency, and sustainable extraction practices. Regular monitoring ensures safe use and prevents contamination, while efficient irrigation techniques maximize yield. Proper maintenance is crucial for long-term sustainability and effective management across all sectors. The proposed research is intended to identify the factors of bore well to compute its efficiency.

4. RESULTS AND DISCUSSION

This division represents and analyses the outcomes of proposed method that classifies the ground water quality with proposed optimized Gradient model.

4.1. Performance Metrics

The metrics are predominantly utilized for observing the effectiveness of the respective DL model by applying various metrics. They are recall, Precision, Accuracy and F1 score.

1. Recall Metric

Recall is utilized to examine the data percentage which is correctly detected in the respective model. The recall formula is defined in the following equation (2),

$$\text{Rec} = \frac{\text{True\_Pos}}{\text{False\_Neg+True\_Pos}} \tag{2}$$

2. Accuracy Metric

Accuracy is an important metric used to analyse the number of estimates which are approximately correct in the present model. The accuracy formula is illustrated in equation (3),

$$\text{Accu} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{3}$$

FP, TN, TP, FN Are False Positive, True Negative, True Positive and False Negative.

3. F1-Score Metric

F1-score is used to analyse the predictions in the present model that are made for positive class. The f1 score formula is mentioned in equation (4),

$$\text{F1} = 2 * \frac{\text{Recall*Precision}}{\text{Recall+Precision}} \tag{4}$$

4. Precision Metric

Precision metric is indicated as the method's covariance unit. It is resulted through suitably predictable cases (True_Pos) to the entire group of cases which have been precisely considered (True_Pos + False_Pos). It comprises repeatability and producibility of the capitals. Equation (5), depicts the formula for precision.

$$\text{Prec} = \frac{\text{True\_Pos}}{\text{False\_Pos+True\_Pos}} \tag{5}$$

4.2 Exploratory Data Analysis

This section provides EDA of the water quality dataset in the respective model. The EDA is applied to visualise and understand the data in the generated dataset. Correlation matrix, statistical technique is utilised to calculate relationship among two variables in the dataset. Figure. 4 signifies the correlation matrix for the features in the proposed model.

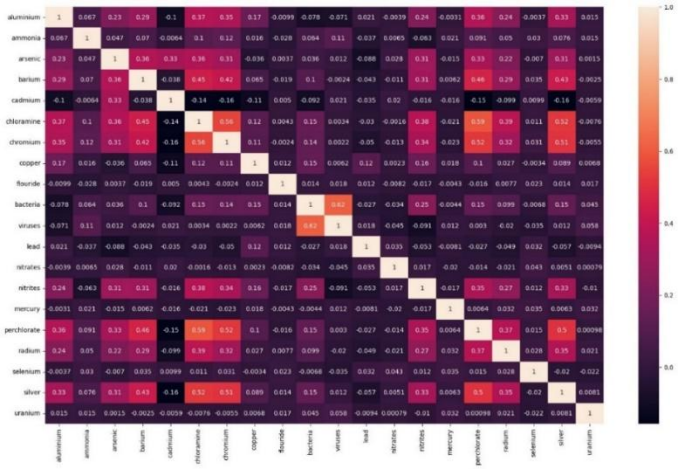

Figure. 4 Correlation Matrix

Figure. 4 illustrates the correlation matrix for proposed ML model. It signifies the relationship among various variables such as aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride,

bacteria, viruses, lead, nitrates, mercury, perchlorate, radium, selenium, silver and uranium. This matrix displays the correlation coefficients among various chemical elements, compounds, and contaminants. The values range from -1 to 1 which is Perfect negative correlation and perfect positive correlation respectively. 0 represents no correlation. Chromium & Chloramine, Perchlorate & Chloramine, Silver & Chloramine and Perchlorate & Silver have strong positive correlation. Cadmium & Arsenic has negative correlation. Water quality is monitored by seeing a high level of one substance may prompt to check for another with which it is strongly correlated. In addition, figure. 5 shows the Counts of Aluminium and Nitrates.

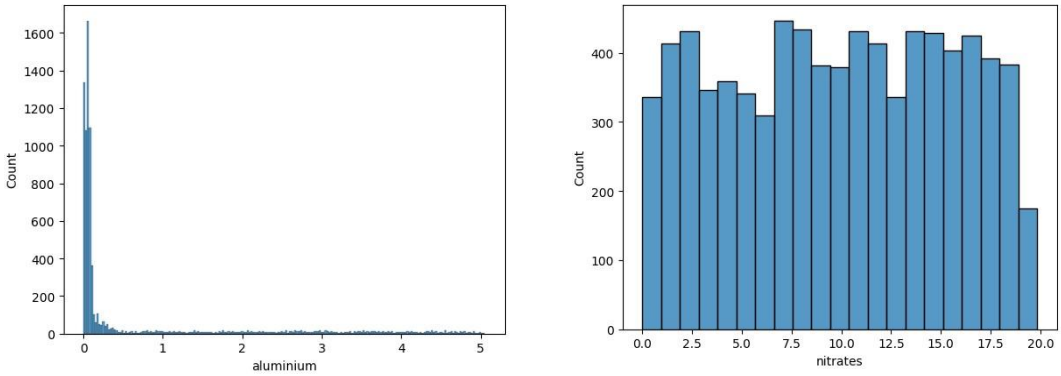

Figure. 5 Counts of Aluminium and Nitrates

From figure. 5, it clears that the Aluminium concentrations are generally low, with most occurrences close to 0. This suggests that in this dataset, high aluminium levels are infrequent, possibly due to effective control measures or natural factors. Secondly, Nitrate concentrations are distributed more evenly across a wider range, suggesting varying levels of nitrate contamination in the dataset. Unlike aluminium, nitrates don't show a strong skew towards lower values, indicating more consistent and varied presence in the samples. Aluminium found in low concentrations in this dataset whereas Nitrates exhibits a broader range of concentrations, indicating more variability in their presence.

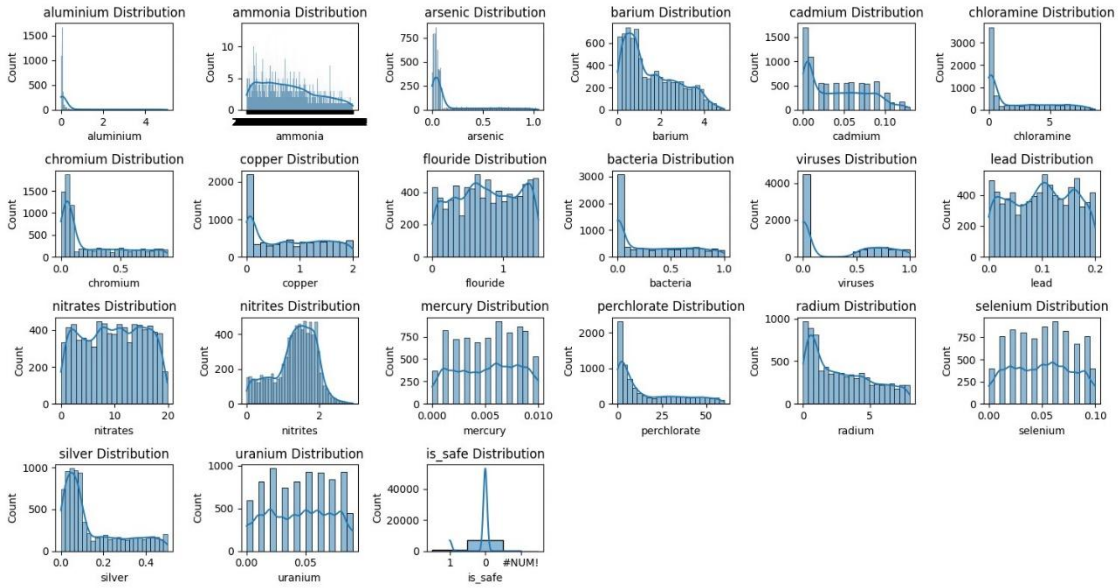

Figure. 6 Counts for Various Chemicals

From figure. 6, it clears that Aluminium, Arsenic, Barium, Cadmium, Chloramine, Chromium, Copper, Perchlorate, and Silver distributions are heavily skewed to the right, meaning most observations are near zero with a few high outliers. Fluoride, Nitrates, Nitrites, Mercury, Selenium, Uranium distributions are more evenly spread or bimodal, indicating a more diverse presence of these elements or compounds across the dataset. Likely, Nitrates and Nitrites show distributions with relatively even counts across different concentration ranges, though Nitrites exhibit a bimodal pattern with peaks around 1 and 2 units. Furthermore, figure. 7 illustrates Principal Component and Its Ranked Features for the proposed ML model.

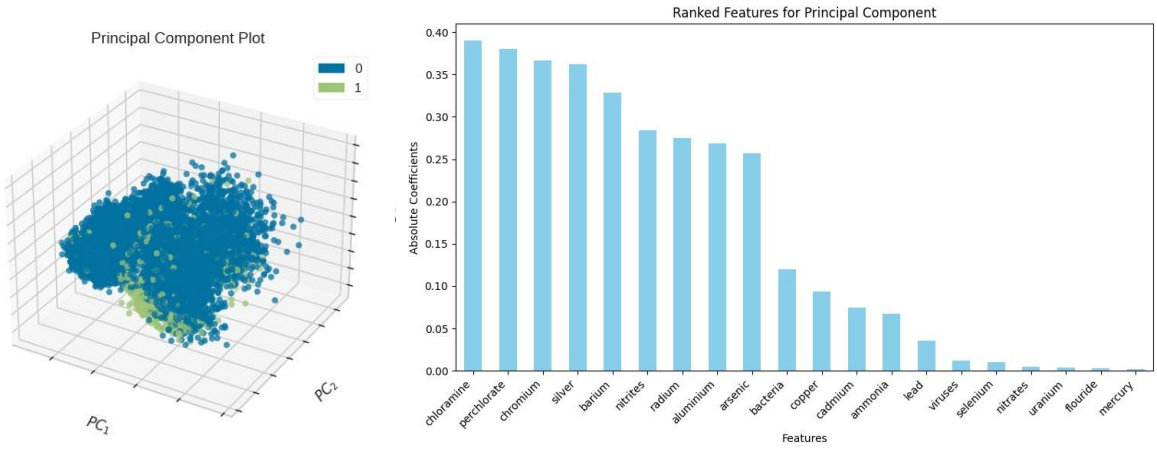

Figure. 7 Principal Component and its Ranked Features

Figure. 7, the 3D scatter plot shows the distribution of samples projected onto three principal components, helping to visualize the clustering or separation of data based on variance. Overlap indicates that the principal components capture some, but not all, of the variance that separates safe from unsafe samples. While the PCA reduces the dimensionality of the data, it shows some clustering of safe vs. unsafe data, though the overlap suggests that more features or additional analysis might be needed to clearly distinguish between the two categories.

4.3 Performance Analysis

The Performance of present DL model is considered utilising several metrics like Precision, Recall, Accuracy and F1-score. Similarly, CM (Confusion Matrix) which was utilised for recognising the presentation. It summarises and visualizes the overall performance of the proposed method which shows how many predictions are 0 & 1 as per the class. In following, figure. 8 illustrates the CM for AdaBoost and Gradient Boosting Models.

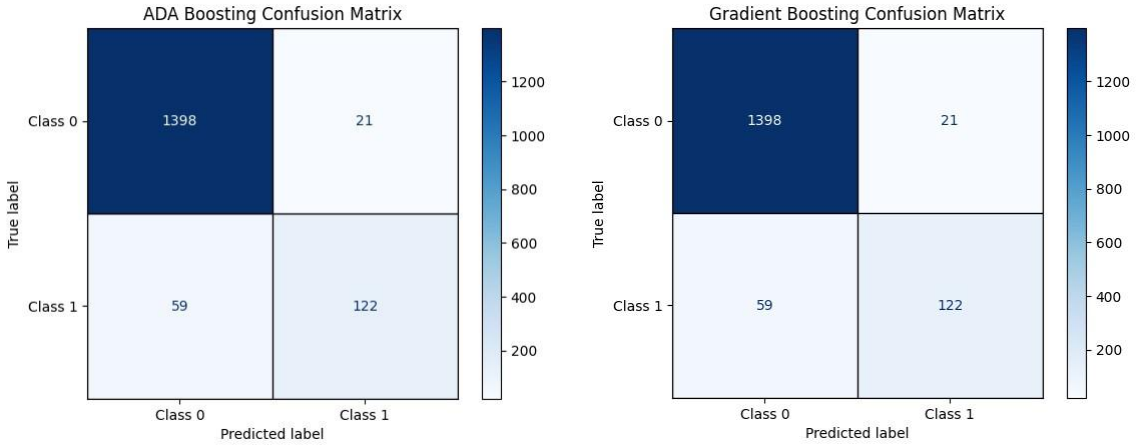

Figure. 8 CM for AdaBoost and Gradient Boosting Models

The figure. 8 illustrates the CM of AdaBoost and Gradient Boosting Models. It deliberates true label and predicted label for water quality dataset. Likewise, figure. 9 shows CM for other two models such as RF and XG Boost models.

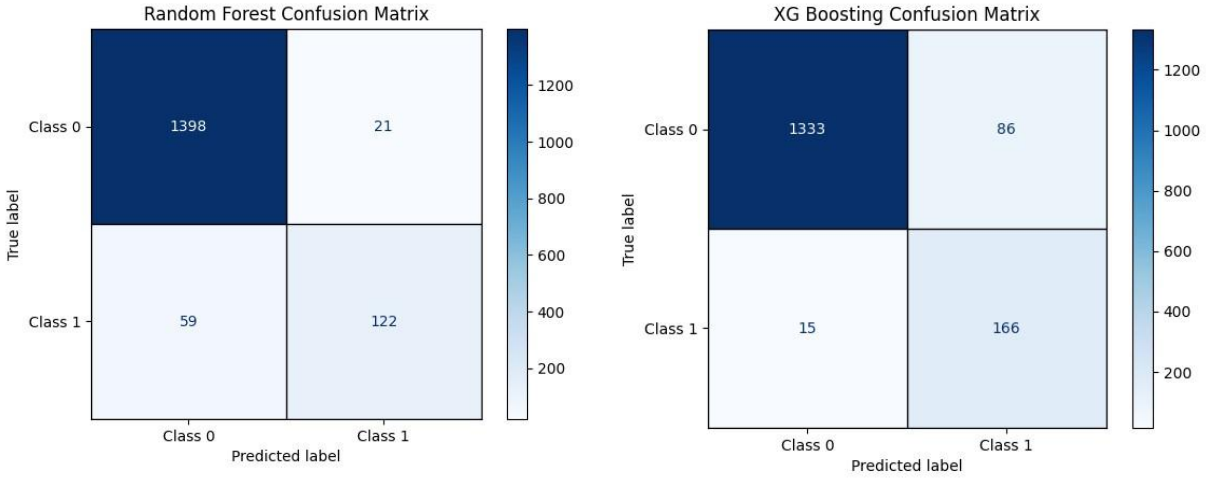

Figure. 9 CM for Random Forest and XG Boosting Models

The figure. 9 illustrates the CM of Random Forest and XG Boosting Models. It shows true label and predicted label for water quality dataset. Here, Class 1 is and 0 is not safe. Likely, figure. 10 illustrates the CM for proposed ML model.

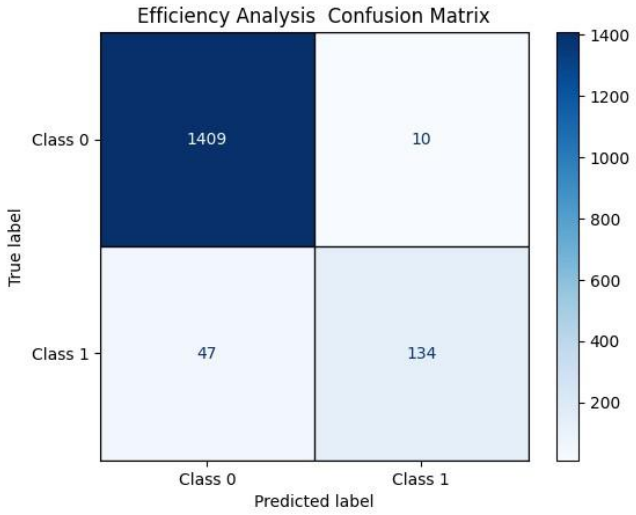

Figure. 10 CM for Proposed Optimized Gradient Model

From figure. 10, it clears that the proposed ML model attained better performance than other models such as AdaBoost, gradient Boosting, RF and XG Boost models. Concomitantly, the table 1 depicts proposed results for bore well water.

Table 1. Proposed Results for Bore well Factors

Factors	Domestic	Agricultural	Industrial	Irrigation
PO_4^{-}	0.749	0.792	0.799	0.725
Depth	0.814	0.792	0.747	0.652
Temp	0.621	0.664	0.703	0.632
SO_4^{2-}	0.703	0.574	0.342	0.47
K^+	0.492	0.239	0.339	0.252
Ca^{2+}	0.502	0.269	0.363	0.252
Na^+	0.587	0.264	0.269	0.432
EC	0.264	0.232	0.229	0.195

HCO3-	0.369	0.294	0.295	0.253
Cl^-	0.253	0.194	0.19	0.286
Mg2+	0.045	0.032	0.009	0.032
Salinity	0.289	0.052	0.035	0.108
pH	0.079	0.067	0.059	0.065
AS	0.087	0.075	0.035	0.057
F-	0.987	0.943	0.933	0.872

The table 1 signifies various usage of bore well water like domestic, agricultural, irrigation and industrial use. Domestic factor attained 0.749, 0.814, 0.621, 0.703, 0.492, 0.502, 0.587, 0.264, 0.369, 0.253, 0.045, 0.289, 0.079, 0.087 and 0.987 and agricultural factor attained 0.792, 0.792, 0.664, 0.574, 0.239, 0.269, 0.264, 0.232, 0.294, 0.194, 0.032, 0.052, 0.067, 0.075 and 0.943 of PO_4^-, Depth, Temp, SO42-, K+, Ca2+, Na+, EC, HCO3-, Cl^-, Mg2+, Salinity, pH, AS and F- respectively.

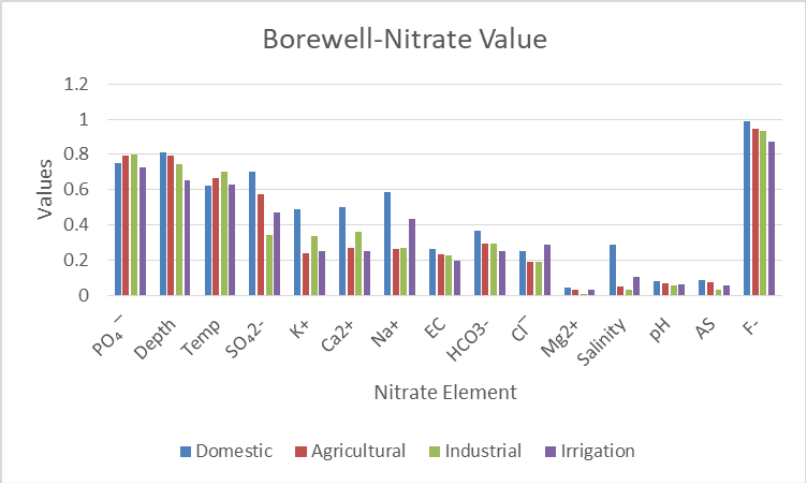

Figure. 11 Graphical Representation of Nitrate Values in Bore well Water

Figure. 11 illustrates about graphical representation of bore well water nitrate values. 0.799, 0.747, 0.703, 0.342, 0.339, 0.363, 0.269, 0.229, 0.295, 0.19, 0.009, 0.035, 0.059, 0.035 and 0.933 and irrigation factor attained 0.725, 0.652, 0.632, 0.47,0.252, 0.252, 0.432, 0.195, 0.253, 0.286, 0.032, 0.108, 0.065, 0.057 and 0.872 of PO_4^-, Depth, Temp, SO42-, K+, Ca2+, Na+, EC, HCO3-, Cl^-, Mg2+, Salinity, pH, AS and F- respectively.

4.4 Comparative Analysis

The section exemplifies both internal and external comparative analysis of the projected DL model in accordance with various performance metrics. The table 2 compares the factors of water quality in all three phases.

Table.2 Comparison of Phase 1, Phase 2 and Phase 3

Factors	Phase-1	Phase-2	Phase-3
PO_4^-	0.689	0.752	0.853
Depth	0.654	0.754	0.824
Temp	0.562	0.623	0.729
SO42-	0.446	0.534	0.632
K+	0.261	0.195	0.375
Ca2+	0.237	0.259	0.369
Na+	0.154	0.168	0.251
EC	0.165	0.179	0.183
HCO3-	0.154	0.261	0.273

Cl^-	0.144	0.158	0.164
Mg2+	0.054	0.036	0.039
Salinity	0.013	0.041	0.049
pH	0.029	0.039	0.042
AS	0.0009	0.063	0.069
F-	0.953	0.972	0.981

Table 1 depicts the water qualities factors comparison. It shows that the present model attained better results than previous phases 1 and phase 2. Phase 3 attained 0.853, 0.824, 0.729, 0.632, 0.375, 0.369, 0.251, 0.183, 0.273, 0.164, 0.039, 0.049, 0.042, 0.069 and 0.981 of PO_4^-, Depth, Temp, SO42-, K+, Ca2+, Na+, EC, HCO3-, Cl^-, Mg2+, Salinity, pH, AS and F- respectively which are higher than phase 1 and phase 2 results.

Internal Comparison

This section shows the internal comparison performance. The table.3 and figure. 11 signifies the comparison analysis of proposed Optimized Gradient model.

Table.3 Comparative Analysis of Optimizing Gradient Boosting

Model	Accu	Prec	Rec	F1
Gradient Boosting Model	0.95	0.85	0.89	0.91
XG Boost Model	0.94	0.89	0.92	0.9
Ada Boost Model	0.93	0.87	0.9	0.92
Random Forest Model	0.94	0.92	0.89	0.88
Optimizing Gradient Boosting with Focal Loss	**0.96**	**0.93**	**0.94**	**0.93**

From table 2, the proposed model attains best performance in the following metrics of accuracy, precision, recall and F1-Score with 0.96, 0.93, 0.94 and 0.93 respectively. The figure 11 represent the performance of other models.

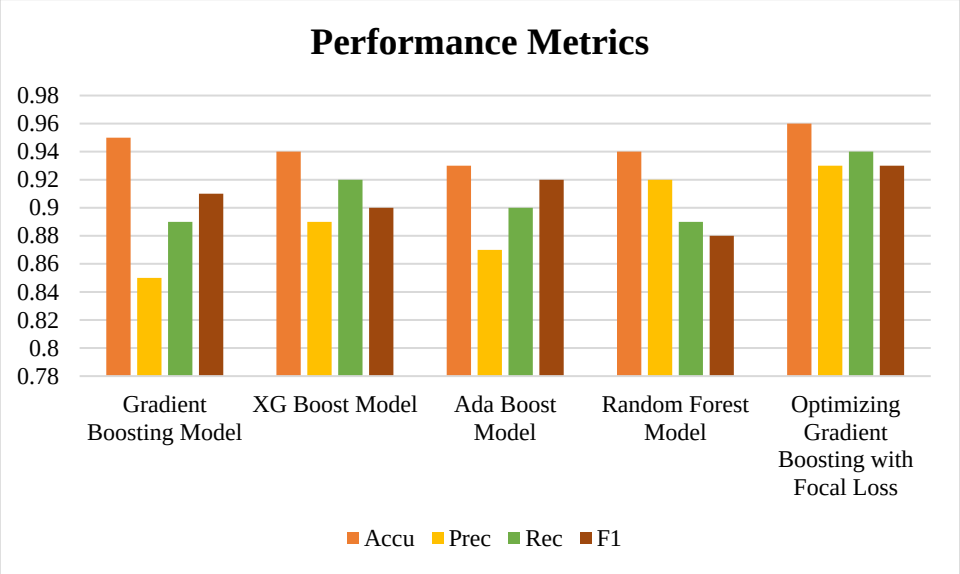

Figure. 11 Comparative Analysis of Optimizing Gradient Boosting

Figure. 11 depicts the comparison of present model by other existing models. The proposed model is compared with Gradient Boosting model (0.95, 0.85, 0.89 and 0.91), XG Boost model (0.94, 0.89, 0.92 and 0.9), Ada Boost model (0.93, 0.87, 0.9 and 0.92) and Random Forest model (0.94, 0.92, 0.89 and 0.88).

External Comparison

This section describes the external comparative analysis of evaluation metrics of proposed model with existing research. The table. 4 presents comparison performance of Proposed Method with conventional research.

Table.4 Comparative Analysis of Optimizing Gradient Boosting

Model	Accu	Prec	Rec	F1
LR (Logistic Regression)	0.9	0.8	0.63	0.67
DT	0.87	0.88	0.88	0.87
RF (Random Forest)	0.93	0.89	0.82	0.87
SVM	0.91	0.82	0.89	0.88
ADA	0.93	0.88	0.74	0.79
ETC	0.92	0.92	0.65	0.7
Optimizing Gradient Boosting with Focal Loss	**0.96**	**0.93**	**0.94**	**0.93**

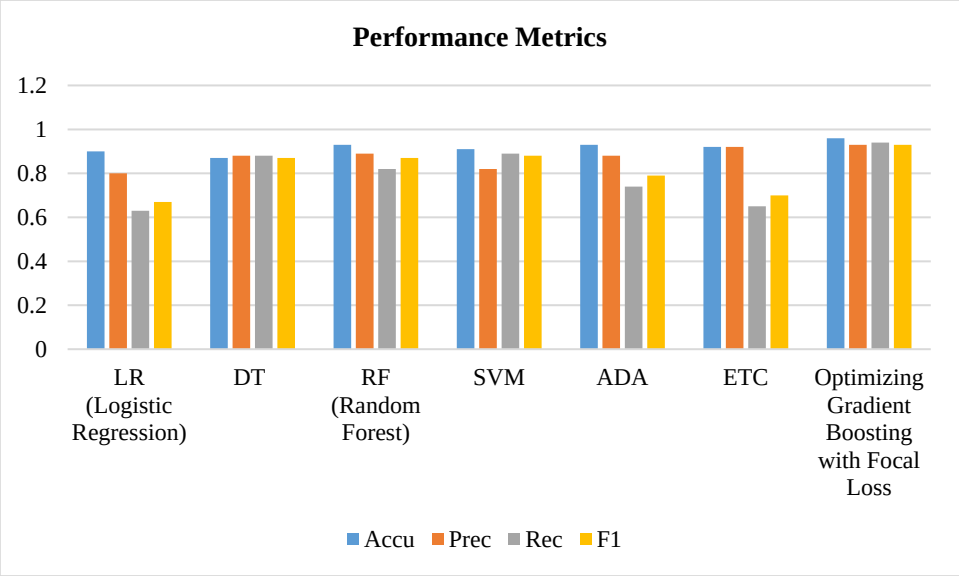

Figure. 12 Comparison of Optimizing Gradient Boosting

From table 3, the proposed model is compared with some existing approaches are LR (0.9, 0.8, 0.63, 0.67), DT (0.87, 0.88, 0.88, 0.87), RF (0.93, 0.89, 0.82, 0.87), SVM (0.91, 0.82, 0.89, 0.88), ADA (0.93, 0.88, 0.74, 0.79) and ETC (0.92, 0.92, 0.65, 0.7). Figure. 12 depicts the graphical representation of the table. 4.

Table.5 Comparative Analysis of Optimizing Gradient Boosting

Model	Accu	Prec	Rec	F1
NB (Naive Bayes)	0.83	0.8	0.94	0.91
RF	0.94	0.88	0.93	0.87
Gradient Boosting	0.93	0.95	0.93	0.87
KNN	0.9	0.92	0.92	0.88

LR	0.9	0.91	0.91	0.91
DT	0.91	0.92	0.93	0.92
Optimizing Gradient Boosting with Focal Loss	**0.96**	**0.93**	**0.94**	**0.93**

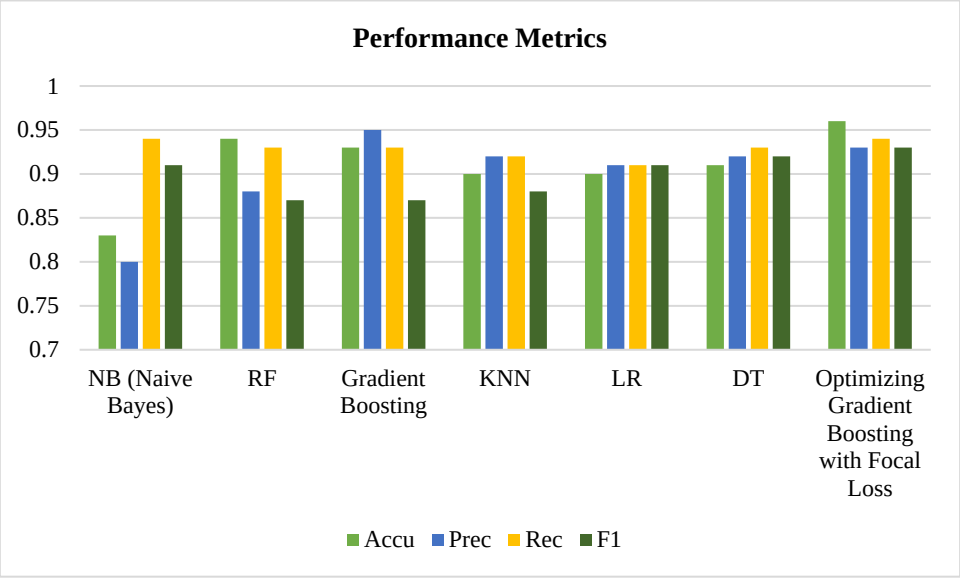

Figure. 13 Comparative Analysis of Optimizing Gradient Boosting

From table 5 and figure. 13, the present model is compared with prevailing techniques are NB (0.83, 0.8, 0.94, 0.91), RF (0.93, 0.89, 0.82, 0.87), Gradient Boosting model (0.95, 0.85, 0.89, 0.91), KNN (0.9, 0.92, 0.92, 0.88), LR (0.9, 0.8, 0.63, 0.67) and DT (0.87, 0.88, 0.88, 0.87). Finally, the proposed model obtains best result when compare to other models. The graphical representation of table 3 shown in the figure 13.

5. CONCLUSION

Globally, groundwater is a significant source of consumption water for countless individuals. Calculating its pureness is a supreme to recognising dangerous pollutants such as nitrates, microbes and heavy metals. Therefore, persistent evaluation and monitoring are necessary in mitigating the risks interconnected with the contamination of ground water and ensuring safety for people. Hence, an effective Groundwater quality classification is significant to ensure the environment security. However, providing accuracy by human professionals is considered to be time-taking and inadequate. To solve this challenge, the present research of Optimized gradient with effective analysis in trees with focal loss model is employed to avoid limitations and improves the detection accurateness for Groundwater quality classification. The process of feature optimization utilised PCA Optimized model. The proposed research utilised water quality dataset to demonstrate the effectiveness. In water quality dataset, the proposed research attained 0.96, 0.93, 0.94 and 0.93 of accuracy, precision, recall and F1score respectively. The internal comparison of proposed model Gradient Boost attained 0.95, 0.85, 0.89 and 0.91, XG Boost Model attained 0.94, 0.89, 0.92 and 0.9, AdaBoost attained 0.93, 0.87, 0.9 and 0.92. Random Forest attained 0.94, 0.92, 0.89 and 0.88 of accuracy, precision, recall and F1 score respectively. Constantly, results from the comparative performance indicated that the proposed ML model has overtook the prevailing researches. In future, the present approach could be applied on various water quality datasets for prediction purpose in order to enhance the entire environmental and public health security. It could assists in further researches of effective groundwater quality detection models.

REFERENCES

[1] N. Chandel, S. K. Gupta, A. K. J. J. o. M. Ravi, and Environment, "Ground Water Quality Analysis using Machine Learning Techniques: a Critical Appraisal," vol. 15, no. 2, pp. 419-426, 2024.

[2] C. R. Das, S. J. E. S. Das, and P. Research, "Coastal groundwater quality prediction using objective-weighted WQI and machine learning approach," vol. 31, no. 13, pp. 19439-19457, 2024.
[3] A. Jose, S. J. W. P. Yasala, and Technology, "Machine learning-based ensemble model for groundwater quality prediction: A case study," vol. 19, no. 6, pp. 2364-2375, 2024.
[4] H. Raheja, A. Goel, and M. J. I. J. o. H. E. Pal, "A novel approach for prediction of groundwater quality using gradient boosting-based algorithms," pp. 1-12, 2024.
[5] K. B. W. Boo, A. El-Shafie, F. Othman, M. M. H. Khan, A. H. Birima, and A. N. J. W. R. Ahmed, "Groundwater level forecasting with machine learning models: A review," p. 121249, 2024.
[6] M. F. Allawi, Y. Al-Ani, A. D. Jalal, Z. M. Ismael, M. Sherif, and A. J. E. A. o. C. F. M. El-Shafie, "Groundwater quality parameters prediction based on data-driven models," vol. 18, no. 1, p. 2364749, 2024.
[7] F. Ding *et al.*, "Optimization of water quality index models using machine learning approaches," vol. 243, p. 120337, 2023.
[8] H. Ghosh, M. A. Tusher, I. S. Rahat, S. Khasim, and S. N. Mohanty, "Water quality assessment through predictive machine learning," in *International Conference on Intelligent Computing and Networking*, 2023, pp. 77-88: Springer.
[9] M. Y. Shams *et al.*, "Water quality prediction using machine learning models based on grid search method," vol. 83, no. 12, pp. 35307-35334, 2024.
[10] M. G. Uddin, S. Nash, A. Rahman, and A. I. J. W. R. Olbert, "A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment," vol. 219, p. 118532, 2022.
[11] F. J. I. J. o. C. Ahmad Musleh and D. Systems, "A Comprehensive Comparative Study of Machine Learning Algorithms for Water Potability Classification," vol. 15, no. 1, pp. 1189-1200, 2024.
[12] R. I. Singh and U. K. Lilhore, "A survey of machine learning models for water quality prediction," in *2023 8th International Conference on Communication and Electronics Systems (ICCES)*, 2023, pp. 1069-1074: IEEE.
[13] N. K. Ravi *et al.*, "Application of water quality index (WQI) and statistical techniques to assess water quality for drinking, irrigation, and industrial purposes of the Ghaghara River, India," vol. 6, p. 100049, 2023.
[14] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, M. K. J. J. o. K. S. U.-C. Nasir, and I. Sciences, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," vol. 34, no. 8, pp. 4773-4781, 2022.
[15] A. Yogeshwari, J. Anubama, M. J. M. M. Jenitha, M. C. Geetha, M. D. Ramalakshmi, and B. J. J. o. S. i. F. S. Pavithra, "Water Quality Prediction Using Catboost Classifier Algorithm," vol. 10, no. 4S, pp. 2850-2855, 2023.
[16] N. Nasir *et al.*, "Water quality classification using machine learning algorithms," vol. 48, p. 102920, 2022.
[17] S. Kouadri, C. B. Pande, B. Panneerselvam, K. N. Moharir, A. J. E. S. Elbeltagi, and P. Research, "Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models," pp. 1-25, 2022.
[18] H. Raheja, A. Goel, M. J. W. P. Pal, and Technology, "Prediction of groundwater quality indices using machine learning algorithms," vol. 17, no. 1, pp. 336-351, 2022.
[19] J. Huan, H. Li, F. Wu, and W. J. A. E. Cao, "Design of water quality monitoring system for aquaculture ponds based on NB-IoT," vol. 90, p. 102088, 2020.
[20] F. Akhter, H. R. Siddiquei, M. E. E. Alahi, K. P. Jayasundera, and S. C. J. I. I. o. T. J. Mukhopadhyay, "An IoT-enabled portable water quality monitoring system with MWCNT/PDMS multifunctional sensor for agricultural applications," vol. 9, no. 16, pp. 14307-14316, 2022.
[21] S.-S. Baek, J. Pyo, and J. A. J. W. Chun, "Prediction of water level and water quality using a CNN-LSTM combined deep learning approach," vol. 12, no. 12, p. 3399, 2020.
[22] R. Barzegar, M. T. Aalami, J. J. S. E. R. Adamowski, and R. Assessment, "Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model," vol. 34, no. 2, pp. 415-433, 2020.
[23] S. Pasika and S. T. J. H. Gandla, "Smart water quality monitoring system with cost-effective using IoT," vol. 6, no. 7, 2020.
[24] A. El Bilali, A. Taleb, and Y. J. A. W. M. Brouziyne, "Groundwater quality forecasting using machine learning algorithms for irrigation purposes," vol. 245, p. 106625, 2021.
[25] M. Hmoud Al-Adhaileh and F. J. S. Waselallah Alsaade, "Modelling and prediction of water quality by using artificial intelligence," vol. 13, no. 8, p. 4259, 2021.

[26] M. M. Hassan *et al.*, "Efficient prediction of water quality index (WQI) using machine learning algorithms," vol. 1, no. 3, pp. 86-97, 2021.

[27] N. H. A. Malek, W. F. Wan Yaacob, S. A. Md Nasir, and N. J. W. Shaadan, "Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques," vol. 14, no. 7, p. 1067, 2022.

[28] S. Khullar, N. J. E. S. Singh, and P. Research, "Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation," vol. 29, no. 9, pp. 12875-12889, 2022.

[29] F. Rustam *et al.*, "An artificial neural network model for water quality and water consumption prediction," vol. 14, no. 21, p. 3359, 2022.

[30] F. Ashfaq, U. J. J. o. C. Aman, and I. Systems, "Prediction Of Water Quality Using Effective Machine Learning Techniques," vol. 2, no. 1, 2024.